The metaepigenomic structure of purified human stem cell populations is defined at *cis*-regulatory sequences. N. Ari Wijetunga, Fabien Delahaye *et al.*

Table of contents:

Illustrations

OVERVIEW

An overview of the data analysis performed in the study is provided in **Figure S1**, below.

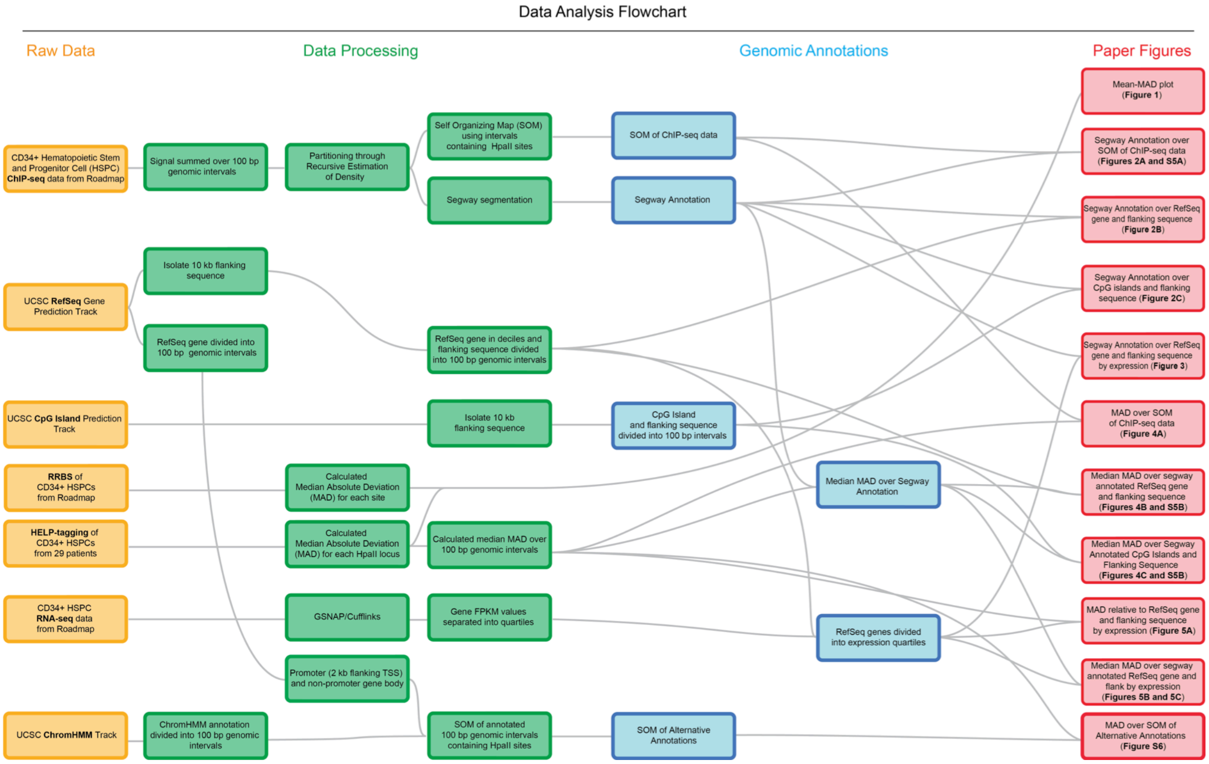
Data Analysis Flowchart
Raw Data
Data Processing
Genomic Annotations
Paper Figures
CD34+ Hematopoietic Stem and Progenitor Cell (HSPC) ChIP-seq data from Roadmap
Signal summed over 100 bp genomic intervals
Partitioning through Recursive Estimation of Density
Self Organizing Map (SOM) using intervals containing HpaII sites
Segway segmentation
SOM of ChIP-seq data
Segway Annotation
UCSC RefSeq Gene Prediction Track
Isolate 10 kb flanking sequence
RefSeq gene divided into 100 bp genomic intervals
RefSeq gene in deciles and flanking sequence divided into 100 bp genomic intervals
UCSC CpG Island Prediction Track
Isolate 10 kb flanking sequence
CpG Island and flanking sequence divided into 100 bp intervals
RRBS of CD34+ HSPCs from Roadmap
Calculated Median Absolute Deviation (MAD) for each site
HELP-tagging of CD34+ HSPCs from 29 patients
Calculated Median Absolute Deviation (MAD) for each HpaII locus
Calculated median MAD over 100 bp genomic intervals
Median MAD over Segway Annotation
CD34+ HSPC RNA-seq data from Roadmap
GSNAP/Cufflinks
Gene FPKM values separated into quartiles
RefSeq genes divided into expression quartiles
Promoter (2 kb flanking TSS) and non-promoter gene body
UCSC ChromHMM Track
ChromHMM annotation divided into 100 bp genomic intervals
SOM of annotated 100 bp genomic intervals containing HpaII sites
SOM of Alternative Annotations
Mean-MAD plot (Figure 1)
Segway Annotation over SOM of ChIP-seq data (Figures 2A and S5A)
Segway Annotation over RefSeq gene and flanking sequence (Figure 2B)
Segway Annotation over CpG islands and flanking sequence (Figure 2C)
Segway Annotation over RefSeq gene and flanking sequence by expression (Figure 3)
MAD over SOM of ChIP-seq data (Figure 4A)
Median MAD over segway annotated RefSeq gene and flanking sequence (Figures 4B and S5B)
Median MAD over Segway Annotated CpG Islands and Flanking Sequence (Figures 4C and S5B)
MAD relative to RefSeq gene and flanking sequence by expression (Figure 5A)
Median MAD over segway annotated RefSeq gene and flank by expression (Figures 5B and 5C)
MAD over SOM of Alternative Annotations (Figure S6)

FIGURE S1: To help guide the reader, we show how we proceeded from Raw Data on the left through the Data Processing steps to generate the Genomic Annotations and ultimately the Figures presented in this report.

Methods: molecular assays

Sample Collection

Cord blood from neonates was the source of material for this study. Biological samples and clinical information were collected (n=37) from consenting women who delivered healthy babies with appropriate growth (matched for gestational age at delivery, ethnicity and gender). Both birth weight and Ponderal index (a measurement of neonatal weight relative to length [1]) were used to identify these normal subjects. Pertinent data on maternal medical history, pregnancy complications and neonatal hospital stay were recorded. (**Table S1**)

TABLE S1: Clinical cohort characteristics (mean ± standard deviation).

	Cohort 1 (n=29)	**Cohort 2 (n=8)**
Gestational age, weeks	39.3 ±0.2	39.2 ±0.4
Birthweight, g	3,110.2 ±51.5	3,206.9 ±164.3
Ponderal index, g/cm^3	2.8 ±0.03	2.8 ±0.07
% Male	50	50
Maternal age, years	26.4 ±0.9	32.9 ±1.8
Pre-pregnancy BMI, kg/m^2	27.2 ±1.4	27.8 ±2.2
Weight gain, pounds	25.7 ±2.8	21.2 ±2.6

Isolation of CD34+ hematopoietic stem and progenitor cells (HSPCs)

CD34+ cells, which constitute approximately 1% of nucleated blood cells in umbilical cord blood [2], were isolated from the cord blood specimen using an immunomagnetic separation technique, as previously described [3, 4]. Mononuclear cells were separated by Ficoll-Paque density gradient or using PrepaCyte-WBC following which CD34+ cells were obtained by positive immunomagnetic bead selection, using AutoMACS Separator (Miltenyi Biotech). This resulted in the isolation of cells with ≥95% purity. We cryopreserved the purified cells in 10% DMSO using controlled rate freezing.

HELP-tagging

HELP-tagging was performed as described [5]. Genomic DNA was isolated from the frozen CD34+ HSPCs, digested to completion by either HpaII or MspI, and then the digested DNA was ligated to two custom adapters containing Illumina adapter sequences, an EcoP15I recognition site and the T7 promoter sequence. Using EcoP15I, we isolated sequence tags flanking the sites digested by each enzyme, methylation-sensitive HpaII or methylation-insensitive MspI, followed by massively-parallel sequencing of the resulting libraries (Illumina technology). HpaII profiles were obtained for each sample (n=29), calculating methylation scores using a previously generated MspI human reference, which was also used to determine the degree of technical variability in the assay, using three replicates.

Bisulphite PCR and amplicon massively-parallel sequencing

We bisulphite-converted 200 ng of DNA using the Zymo EZ-96 Methylation-Lightning Kit. After separate PCR amplification of 10 target regions (primers listed in **Table S2**), we pooled the amplicons in equal ratios and generated Illumina libraries using robotic automation (Tecan). In total, 15 libraries were multiplexed on the Illumina Miseq for 250 bp paired end sequencing. Bisulphite conversion efficiency was calculated as the percent conversion of cytosines in a non-CG context (**Table S3**).

TABLE S2: Bisulphite PCR primers used.

	Locus	Forward Primer	Reverse Primer
ACIN1	chr14:23537766-23538115	GGTTTTGATTGGGTAGTTTAGTAGTAG	CCTACACCCACTTCAAAAAACTT
CD34	chr1:208084599-208084916	ATTAATTTGAGTTTTTTTGGTTGGTT	TCCTTACAATAAATTCCTTTTACAAAATT
CD38	chr4:15779968-15780269	TATTTTAGGATGAATTTTAGTGTATTTGA	CCAACCTCTCTCTTACTACCTAACCT
FRMD4A	chr10:14166941-14167213	TAGTATTTGGAGAAGGATAGAAAGATAGA	ACAAACTACCATCAATCAAAAAAAA
GAPDH #1	chr12:6644694-6644878	GTGGTTGTGGTATGGTGTTAAGT	AACATTTACAACCTAACCTTTAAAATC
GAPDH #2	chr12:6646087-6646282	GAGATTTTTTTAAAATTAAGTGGGG	CCCCCTACAAATAAACCTACAAC
HIVEP3	chr1:42204794-42205064	TAAAGTGTGGTAAAGTTTAAGGAGTT	ACTAAACCTCCCTCTCTAACAAATTA
IF116	chr1:158980701-158980938	TTTGGTATAATATTTTATGTGTGTTTAAA	ATCTCCCCCAAAATACTAAAATTACA
PLB1	chr2:28829909-28830030	GTAGAGTTTAGAGTTGTATGGAGGAG	CACCCCAAACCTAAAATCAAACTTAACA
RPL23A	chr17:27051241-27051430	TATTTATTTTAGGGGAGTGTGGATT	AACCTCTCCACCTTCTAAACCTAC

Two loci were tested at the *GAPDH* locus, denoted as #1 and #2.

TABLE S3: Bisulphite conversion efficiencies (C→T in CH context).

Sample	Conversion efficiency (%)
85	99.630
193	99.754
96	99.772
111	99.578
229	99.650
42	99.761
87	99.724
262	99.742
267	99.742
268	99.751
269	99.730
270	99.773
274	99.732
280	99.728
285	99.722
Average	99.719

Methods: analytical approaches

Calculation of variability of DNA methylation between individuals

We calculated variability of DNA methylation between individuals using the Median Absolute Deviation (MAD) value, an approach previously used in cancer outlier profile analysis (COPA) [6] and to define variability methylated regions [7]. Biological variability is not associated with a specific outcome, which makes testing variability at any single position in the genome difficult because the background level of variation in methylation measurements is unknown *a priori*. For this reason, we utilize the MAD value, a centered and scaled approach which is less sensitive to outliers, giving us a more robust variance estimate. MAD was calculated using the *mad* function implemented in the package *stats* in R (2.15.0). DNA methylation measured by HELP-tagging was used to calculate the MAD variability for our 29 samples from babies with normal birth weight and Ponderal index values in our cohort. To determine variability in the CD34+ HSPCs tested using reduced representation bisulphite sequencing (RRBS), we used publicly available RRBS data for 7 adult CD34+ mobilized hematopoietic stem cell samples from the Roadmap in Epigenomics (**Table S4**) and applied the same approach to calculate the MAD value for each individual site tested by RRBS.

TABLE S4: Characteristics of donors of CD34+ HSPCs used by Roadmap in Epigenomics to generate RRBS data.

Donor	**Age**	**Sex**	**Ethnicity**	**GEO accession number**
RO_01536	27 years	Female	Hispanic	GSM772729
RO_01480	27 years	Female	Caucasian	GSM706858
RO_01549	33 years	Female	Caucasian	GSM706857
RO_01508	36 years	Male	Caucasian	GSM706859
RO_01562	42 years	Female	Caucasian	GSM772730
RO_01517	43 years	Male	Caucasian	GSM706839
RO_01520	50 years	Female	Caucasian	GSM772731

GEO: Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo/

RO_01549 (shaded) is the youngest individual for whom RRBS, chromatin and RNA studies were all performed. We therefore used her data for the further analyses on CD34+ HSPC chromatin and transcription, described later in this document.

Amplicon bisulphite sequence alignment, methylation calls, calculation of variability

Sequence reads from the Illumina MiSeq were trimmed for adapter sequences and aligned to the human genome using *bsmap* (Bisulphite Sequencing Mapping Platform) [8] using the default settings, requiring a PHRED score of ≥37 during alignment. Mean coverage of HpaII sites was 6,676-fold. We checked for bisulphite conversion efficiency (C→T in CH contexts) and quantified the percent methylation for each sample at every CpG in the amplicons using the *methratio* tool provided by *bsmap*.

We performed verification studies on 7 samples from the 29 tested using HELP-tagging, and for validation we added 8 new CD34+ HSPC samples from babies of normal birth weight. Using the methylation percentage at each assayed HpaII locus, we calculated the MAD of methylation for both validation and verification data sets.

We plotted the results in the context of the frequency distribution of the MAD from our 29 subjects tested using HELP-tagging (**Figure S2**). We show the technical variability from MspI assays for comparison, and the result of permuting the methylation values by site within individuals and then recalculating the distribution of variability. Both verification and validation sample sets showed a marked increase in DNA methylation variability when we surpass the background technical variability (defined as the 98.5th centile of MspI range of MAD values).

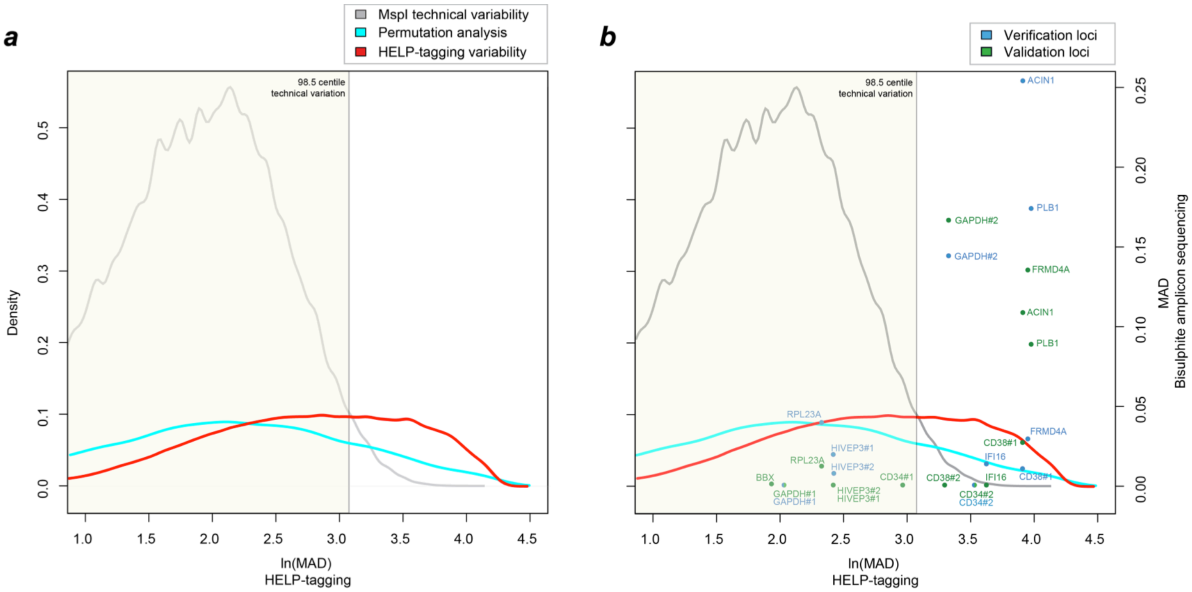

FIGURE S2: (a) The MAD distribution for DNA methylation (measured using HELP-tagging) is shown in red, while the MAD distribution for MspI alone, representing technical variability inherent to the assay, is plotted in grey. Permutation of DNA methylation within each individual was performed to generate the distribution shown in blue. The observed epigenetic variability exceeds both the technical and the expected random variability distributions.

(b) Epigenetic variability calculated as the MAD value for bisulphite amplicon sequencing (using the primers described in **Table S2**) is also plotted. Verification (testing loci from samples already tested using HELP-tagging) and validation (testing new samples for which no genome-wide assays were performed) show that MAD values remain low until the range of technical variability (represented at the 98.5th centile of the MspI distribution) is exceeded.

Common single nucleotide polymorphism (SNP) integrative analysis

The dbSNP build 135 UCSC genome browser track for common SNPs (snp137Common, 11,525,489 SNPs with Minor Allele Frequency ≥1%) was cross-referenced with the tetranucleotide CCGG HpaII site sequences (2,297,198 sites) across the genome. In total, 158,151 HpaII sites directly overlapped a common SNP. Of 1,699,393 HpaII sites for which we measured the MAD values, 1,585,914 did not overlap a common SNP and 113,479 directly overlapped at least one common SNP. The density of the natural log of the MAD values was plotted for HpaII sites not overlapping and overlapping SNPs, respectively (**Figure S3**). The variability in DNA methylation cannot alone be explained by common SNPs overlying HpaII sites. Instead, variability in DNA methylation appears to be indistinguishable at loci with increased potential DNA sequence polymorphism as the remainder of the genome.

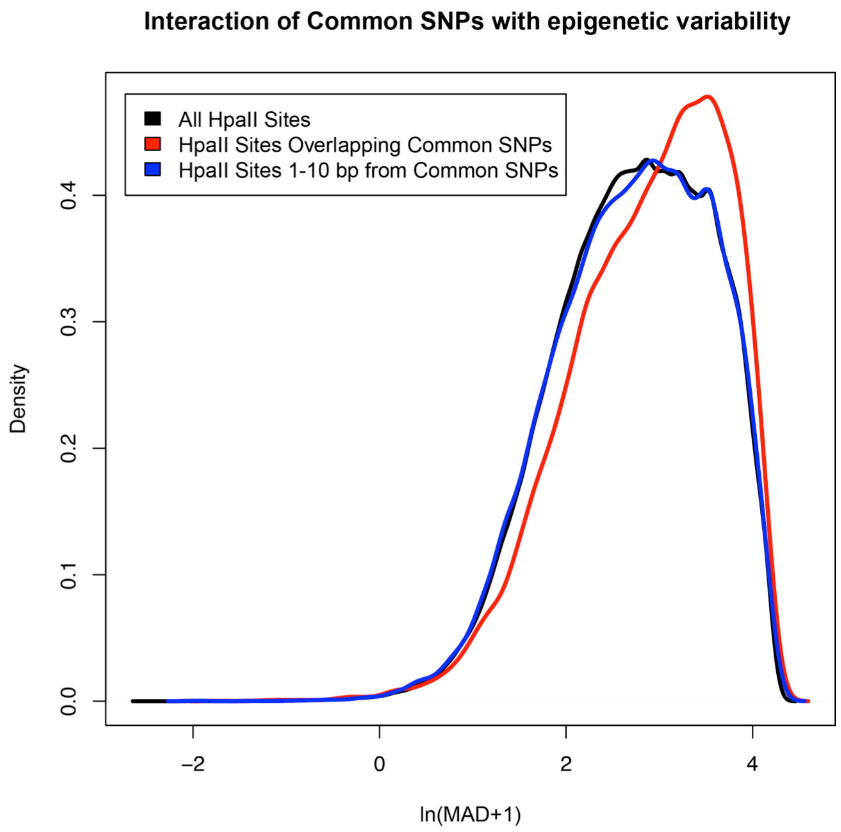

FIGURE S3: MAD distributions of variability of DNA methylation (measured using HELP-tagging) shown for all loci (black) or where a Common SNP (minor allele frequency ≥1%) is found at any one of the nucleotides recognized by HpaII (red, 7.6% of sites tested). The variability at these loci is significantly increased relative to the genome-wide distribution (Kolmogorov–Smirnov test (K–S test) $p<2x10^{-16}$). When the 10 bp immediately flanking the HpaII site overlapping the Common SNP is tested, the distribution of variability reverts to that of the genome as a whole (K-S p=0.1563). We conclude that genetic variability influences DNA methylation readings at loci overlying Common SNPs but that there exists a substantial amount of epigenetic variability in the remaining majority of loci in the genome.

Roadmap chromatin studies

We obtained publicly available chromatin immunoprecipitation followed by massively-parallel sequencing (ChIP-seq) data from the Roadmap in Epigenomics project for CD34+ mobilized HSPCs. From the same individual, a 33 year old, Caucasian female (RO_01549 in **Table S4**), we were able to obtain a measurement of chromatin accessibility (DNase hypersensitivity) as well as ChIP-seq data for six histone modifications (**Table S5**). The raw data provided through http://www.roadmapepigenomics.org/ were minimally preprocessed and available as downloadable wiggle tracks representing density graphs of raw signal from the aligned read density. Quality metrics for the Roadmap preprocessing are shown (**Table S5**).

TABLE S5: Quality metrics for Roadmap in Epigenomics data used in this study.

Regulatory Mark	GEO Accession	Mapped Reads	Findpeaks score	Hotspot score	IROC score	Poisson score
DNase hypersensitivity	GSM530657	-	0.7758	0.6542	0.9981	0.6468
H3K4me3	GSM621439	-	0.6322	0.5962	0.9981	0.6523
H3K4me1	GSM621451	-	0.4931	0.4797	0.9991	0.5749
H3K27me3	GSM706844; GSM621431	15,769,151	0.2382	0.3081	0.9771	0.4494
H3K27ac	GSM772894	12,391,793	0.1533	0.2444	0.9589	0.2649
H3K36me3	GSM706843	18,171,743	0.5987	0.4892		0.7004
H3K9me3	GSM621436	-	0.2508	0.3224		0.5107

Segmentation of data into 100 bp windows

All wiggle tracks were converted to bigwig format using the UCSC Genome Browser utility *wigToBigWig* version 4 (http://hgdownload.cse.ucsc.edu/downloads.html).

Subsequently, the utility available through the UCSC Genome Browser *bigWigAverageOverBed* (http://hgdownload.cse.ucsc.edu/downloads.html) was used to calculate the sum of the ChIP-seq signals over 100 bp genomic intervals spanning the 22 autosomal and 2 sex chromosomes, a resolution which we believe is sufficient to characterize chromatin states, being smaller than an individual nucleosome while usually including no more than 0-1 HpaII sites per window. ChIP-seq signals summed over 30,956,785 intervals were generated and formatted in bedGraph format.

Signal processing approach to define peaks

The Roadmap in Epigenomics data are provided as raw signals and not as defined peaks. To avoid imposing excessive processing on these data, we used as simple and intuitive an approach as possible. The ChIP-seq bedGraph files were log-transformed to exaggerate highly-positively skewed signal density, a prerequisite for the recursive kernel density learning framework for robust foreground object segmentation approach [9]. This image processing technique relies on removing background until the remaining signal is bimodal and approximately Poisson distributed. Gaussian kernel density was estimated using the *density* function in R resulting in multiple modes of signal density that are increasingly smaller. Because signal intensity was derived from a ChIP-seq read count, the measure should be an approximate Poisson distribution, and we aimed to eliminate low signal intensity signal modes iteratively. Using the *turnpoints* function within the *pastecs* library in R, we recursively identified the modes of signal intensity and set signal thresholds based on the maximum mode, which was also generally the leftmost mode. The algorithm ran until at most 2 signal modes remained and the resulting distributions were approximately Poisson distributed. The results are shown in **Figure S4**, with the stepwise approach for the H3K4me1 signal illustrated as an example of the process.

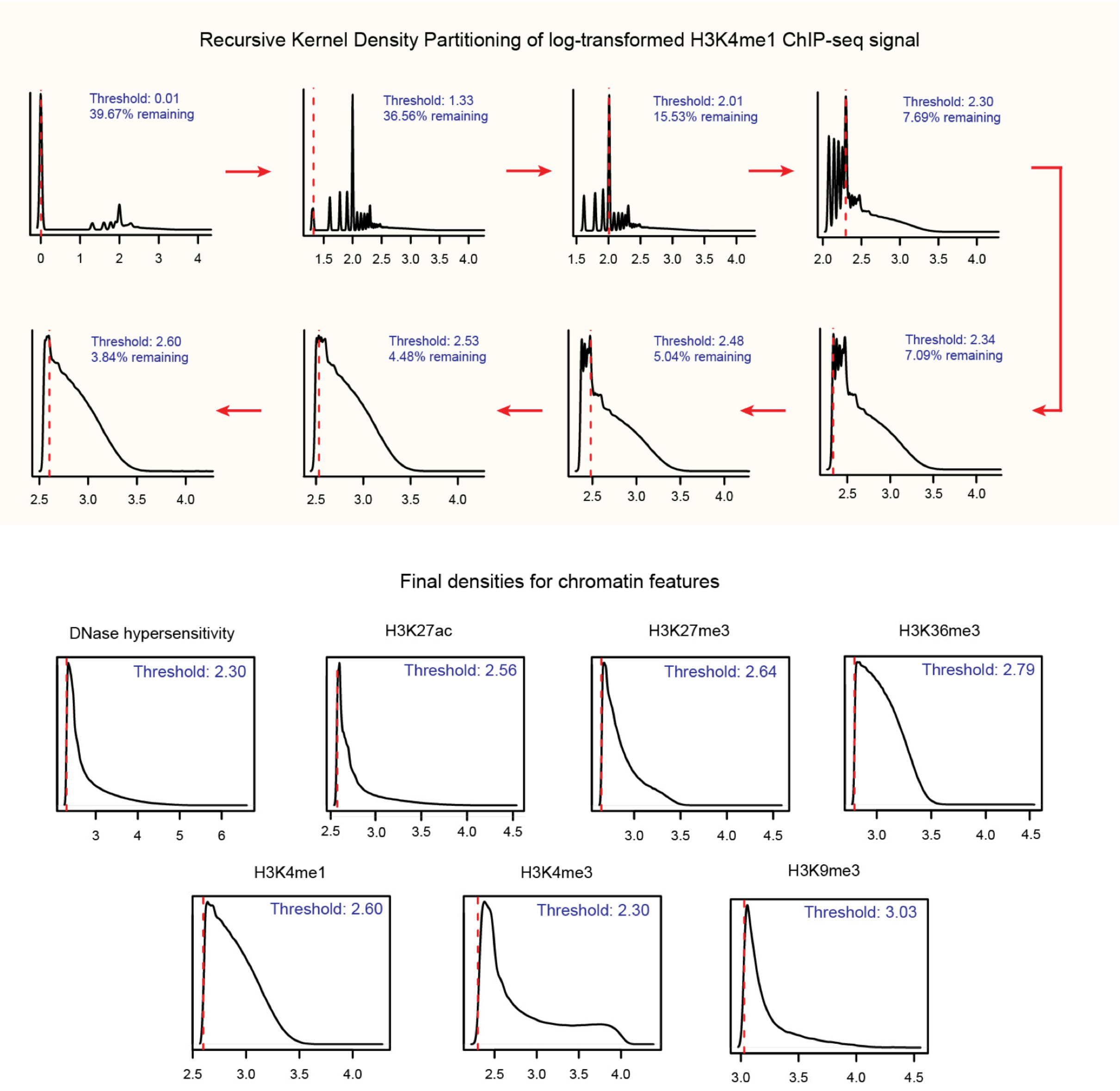

FIGURE S4: When signals from the chromatin features (DNase hypersensitivity, ChIP-seq) are summarized in each 100 bp window in the genome, most windows have little to any signal, as shown on the top left, representing the log-transformed raw H3K4me1 signal. The recursive kernel density partitioning approach finds and removes background, as shown in the series of steps in the upper, shaded panel. When it finds what appears to be a bimodal, approximately Poisson-distributed signal, the segmentation terminates, generating the distributions and thresholds illustrated in the lower panel. Each 100 bp window exceeding this threshold is therefore defined as positive for the chromatin state, the foundation for the analyses that follow.

Self-organising map (SOM) analysis

All SOM analysis [10] was completed using the Java SOMToolbox from Institute of Software Technology and Interactive Systems at the Vienna University of Technology (http://www.ifs.tuwien.ac.at/dm/somtoolbox/). Out of 30,956,785 100 bp genomic intervals, 1,520,684 intervals overlapping 1,696,696 HpaII sites were chosen to reduce the dimensionality of the dataset and greatly reduce the required computation. An input data matrix was created, where rows represent the 1,520,684 vectors defined by genomic intervals and the columns represent the observations for the investigated tracks (*i.e.* processed ChIP-seq signal). In total, two SOMs were constructed, one representing the chromatin states from the Roadmap in Epigenomics, the other the ChromHMM annotation [11], choosing a map size to yield 100-200 interval-vectors per map unit to reduce the required computation while generating maps of sufficient resolution to aid in further analysis. For all SOMs the standard SOM algorithm by Kohonen was employed, using the default SOMToolbox settings of learnrate=0.7 and randomSeed=11. Java code was implemented over approximately 120 hours using a high performance computing (HPC) cluster and 100 GB of virtual memory.

Segway analysis

We performed Segway analysis to define CD34+ HSPC genomic annotations using the Segway genomic segmentation approach [12]. Using the *segway* package, annotations were generated from the 7 processed chromatin state signal bedGraph files. A Segway segmentation of the genome was created by training on chromosome 1, using the results to annotate the whole genome by requesting 7 labels, allowing each chromatin state signal to vary independently from the others. Furthermore, we required at least 1,000 bp segments, a 500 bp ruler and a 500 bp ruler scale, which had the effect of smoothing across the segmentation, which we found to be excessively sensitive to varied signal at the default settings. Segway was completed requesting 10 simultaneous runs, over which maximum likelihood estimations regarding chromatin state were performed. Segway code was implemented using the HPC cluster with the training step taking approximately 72 hours and the identify step taking approximately 2 hours using 40GB of virtual memory.

Contour plots of Variability and Segway features onto SOMs

Enrichment within SOMs was determined by a proportion test, specifically asking whether the observed proportion of a feature within a map unit was significantly greater than the expected proportion given the distribution of Segway features overall. A cutoff of 23.09, the 98.5th centile of overall MAD, was used in order to dichotomize MAD into high and low variable states (for reference, note that ln(23.09) = 3.14, the cutoff shown in **Figure S2**). From calculated proportions of highly variable 100 bp intervals within the SOM units, we performed a proportion test, specifically testing whether the observed proportion of a highly variable interval within a map unit was significantly (alpha=0.05) greater than the expected proportion, given the overall proportion of highly variable intervals. The *MASS* library in R along with the *kde2d* and *contour* functions were used to represent the density of variability-enriched map units with a contour plot over SOMs.

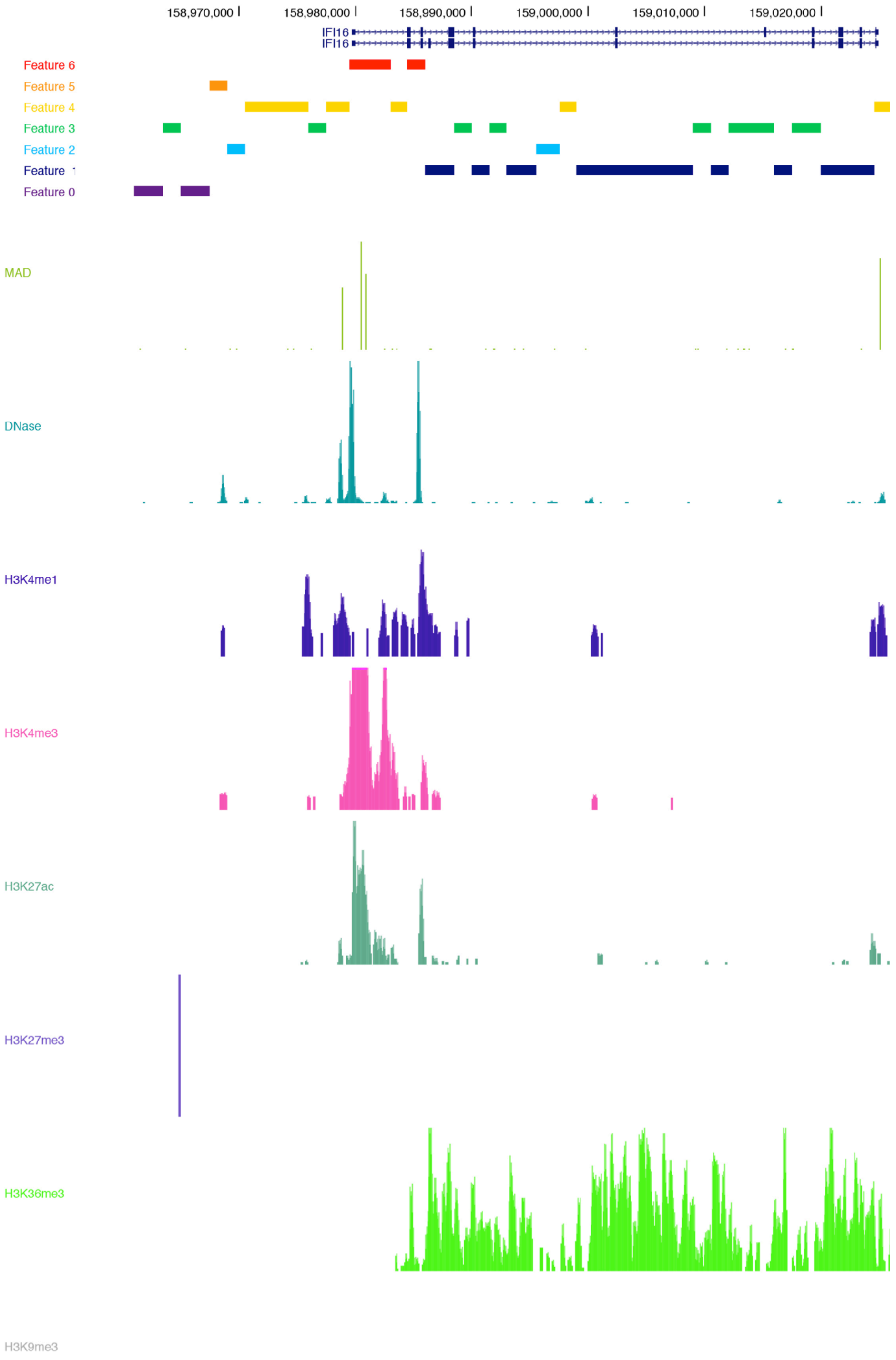

FIGURE S5: Segway features (upper part of panel) derived from chromatin study data (below). The DNA methylation variability track (MAD) shows increased variability at the candidate promoter (Feature 6, red) and enhancer (feature 4, yellow) Segway features.

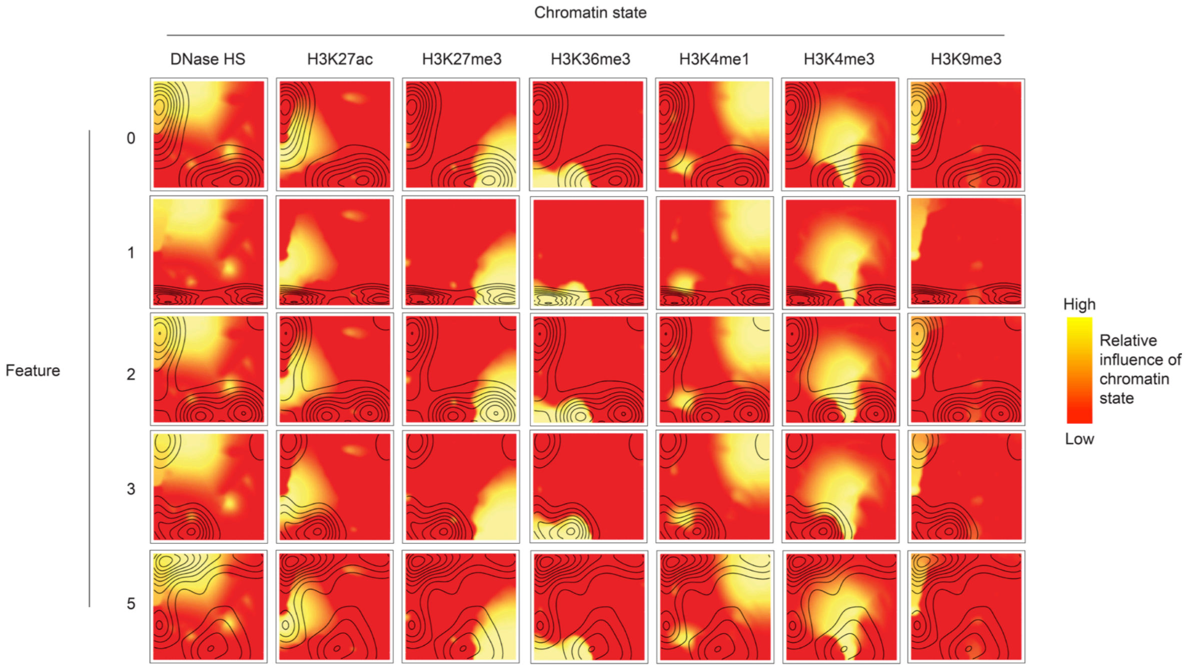

FIGURE S6: SOM analysis of chromatin states, with overlying contour plots for Segway features 0-3 and 5. To interpret the plots, we look for enrichment of certain chromatin states (yellow) at areas with the highest points on the contour plot. For example, Feature 1 shows enrichment of H3K27me3 and H3K36me3, which we interpret to represent transcribed gene bodies.

Meta-plots of Segway features in the contexts of RefSeq genes and CpG islands

In addition to analyzing RefSeq gene and CpG island annotations, we analyzed the 10 kb regions flanking these annotations in order to capture information about flanking regulatory elements. The bodies of RefSeq genes were split into deciles in order to be able to compare genes of varying lengths, and the 10 kb flanking region was separated into 100 bp windows. Gene coordinates were rounded to the nearest 100 bp to ensure that sequences were not represented twice. Similarly, 10 kb regions flanking CpG islands were divided into 100 bp windows. Segway features were divided into 100 bp intervals and matched with 100 bp windows or matched with RefSeq gene deciles, allowing more than one Segway feature to match a particular window or decile. For each 100 bp window or decile, the frequency of each Segway feature was calculated and plotted.

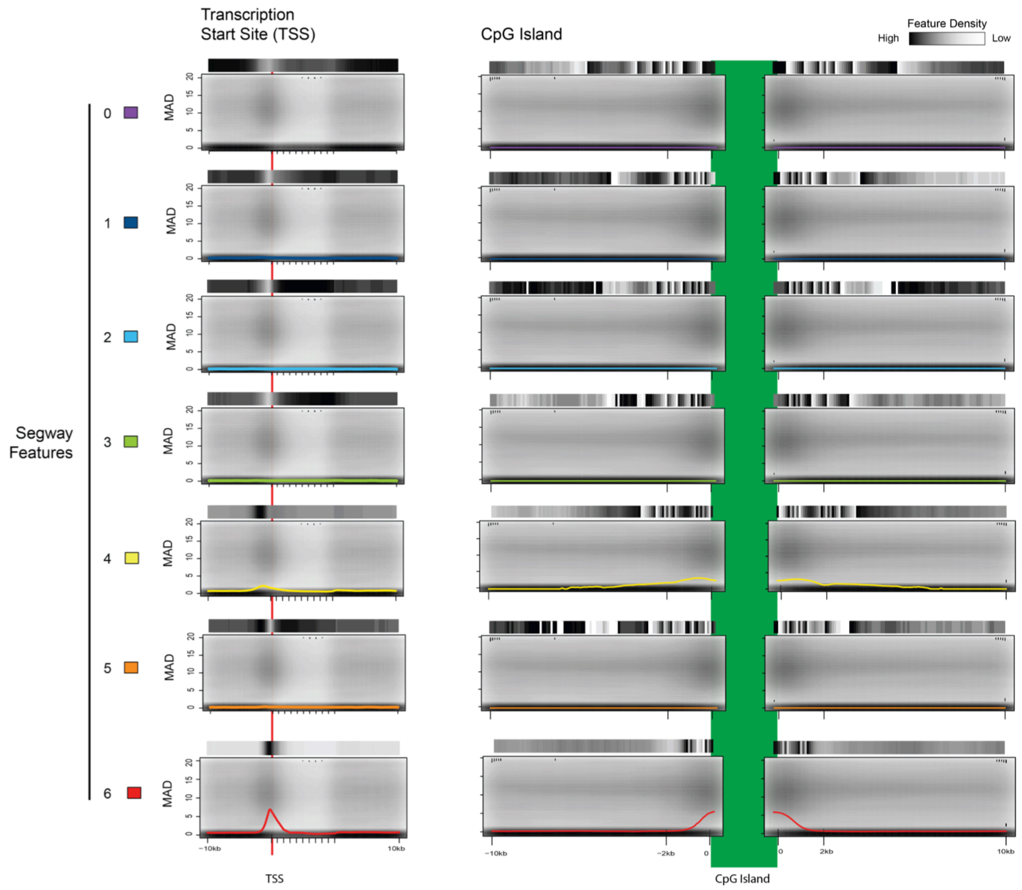

FIGURE S7: Meta-plots of Segway features related to RefSeq gene and CpG island annotations. The darker shading indicates a greater density of the feature in the genomic context shown. No increase of variability of any Segway features other than 4 and 6 is apparent, either with respect to RefSeq genes or CpG islands.

Roadmap transcription data

RNA-seq data processing, gene ranking

CD34+ HSPC RNA-seq data were obtained from individual RO_01549 (**Table S4**), GEO accession number SRA010036, the same individual from whom ChIP-seq and DNase hypersensitivity data were used. Of the 17 available RNA-seq runs, we used SRR453391, corresponding to 16,000,000 reads and 2.4 gigabasepairs of sequence. Reads were quality controlled using *FASTX-Toolkit* v0.0.13, with *fastq_quality_trimmer* trimming nucleotides with quality lower than 3 and removing sequences shorter than 17 bp (http://hannonlab.cshl.edu/fastx_toolkit/commandline.html#fastx_trimmer_usage). Reads were aligned to the human genome using *GSNAP* version 2012-07-20 requesting at most 10 alignments for multiple aligned reads. *SAMtools* v0.1.8 was used to convert the alignment to BAM format.

Stratification by expression quartile of Segway meta-plots on RefSeq genes

The *Cufflinks* v2.02 program *Cuffdiff* v2.0.2 was used to calculate fragments per kilobase per million reads (FPKM) values for RefSeq genes, employing normalization by the upper quartile of the number of fragments mapping to individual loci and default weighting of multiply aligned reads based on the number of alignments. FPKM values were used to separate genes into those that were not expressed (8,963 genes) and those that were expressed by quartile of expression (7,872 genes per quartile). Expression information was then linked to the annotated RefSeq gene body deciles and 10 kb flanking regions, thus allowing the stratification of Segway features overlapping 100 bp windows and gene body deciles by gene expression.

Variability studies

Variability (MAD) tested using a Roadmap chromatin SOM overlay

Each of the seven chromatin state bedGraph tracks was filtered to include only those intervals overlapping HpaII sites. Only 616,114 intervals showed variation in signal between the 7 tracks, and were therefore included in the analysis. Signals were sequentially quantile normalized and unit length normalized, so that signals were comparable depending on the individual distribution of each chromatin state signal. A 79 unit by 79 unit SOM was constructed using the default learnrate=0.7 and randomSeed=11. We selected iterations based on the recommended 5 times the number of observed vectors (3,080,570) iterations to provide sufficient map restructuring given the number of vectors. The output wgt file was parsed to understand the local influences of epigenomic ChIP-seq data within the map, and is directly interpreted as a color weighting of a heat map color scheme (yellow to red scale). The output unit file indicates the vectors associated with a particular map unit, and was parsed to understand enrichment of variability (dichotomized by a cutoff MAD of 23.09) and chromatin state annotations within the map by proportion tests (alpha=0.05).

Variability (MAD) tested by RefSeq gene annotations

Having previously divided Segway features into 100 bp intervals, matching the intervals and features to 100 bp windows of the 10 kb flanking RefSeq genes and within the RefSeq gene body deciles, respectively, we calculated the median MAD for each Segway feature and assigned its variability to the appropriate windows and deciles. For each Segway feature, the median MAD of each window was calculated and scaled by the proportion of the total number of features overlapping the window represented by the feature of interest. The process was repeated for the deciles of RefSeq gene bodies.

CpG island analysis

Likewise, having previously divided Segway features into 100 bp intervals, matching the intervals and features to 100 bp windows within the CpG islands and in the 10 kb flanking region, we calculated the median MAD over the Segway features and assigned variability to the appropriate windows. For each Segway feature, the median MAD over each window was calculated and scaled by the proportion of the total features represented by the feature of interest.

Testing enrichment of Segway features at genomic annotations

We considered the observed proportion of basepairs represented by the intersection of each Segway feature and each genomic feature/annotation compared to the expected proportion that would occur by chance given the respective genomic coverages of each. We then performed a one way proportion test with continuity correction of the observed proportion compared to the expected proportion to calculate the statistical significance of the enrichment at each genomic context. The results are represented as a heat map in **Figure S8**. We observe enrichment of feature 6 at CpG islands and feature 4 at CpG island shores, with enrichment of features 4 and 6 within ±2 kb of RefSeq gene transcription start sites. A ChromHMM annotation of the GM12878 lymphoblastoid cell line was used as a hematopoietic cell comparison with our CD34+ HSPCs, showing feature 6 to enrich at ChromHMM promoter candidates and feature 4 additionally at enhancer candidates, features 1 and 3 at transcribed regions, feature 0 the only one to enrich strongly at ChromHMM-defined heterochromatin, with feature 2 only weakly correlated with ChromHMM-defined annotations for transcribed loci. Although different cell types, the patterns of concordance are reassuring as indicators of the strength of the Segway annotation approach.

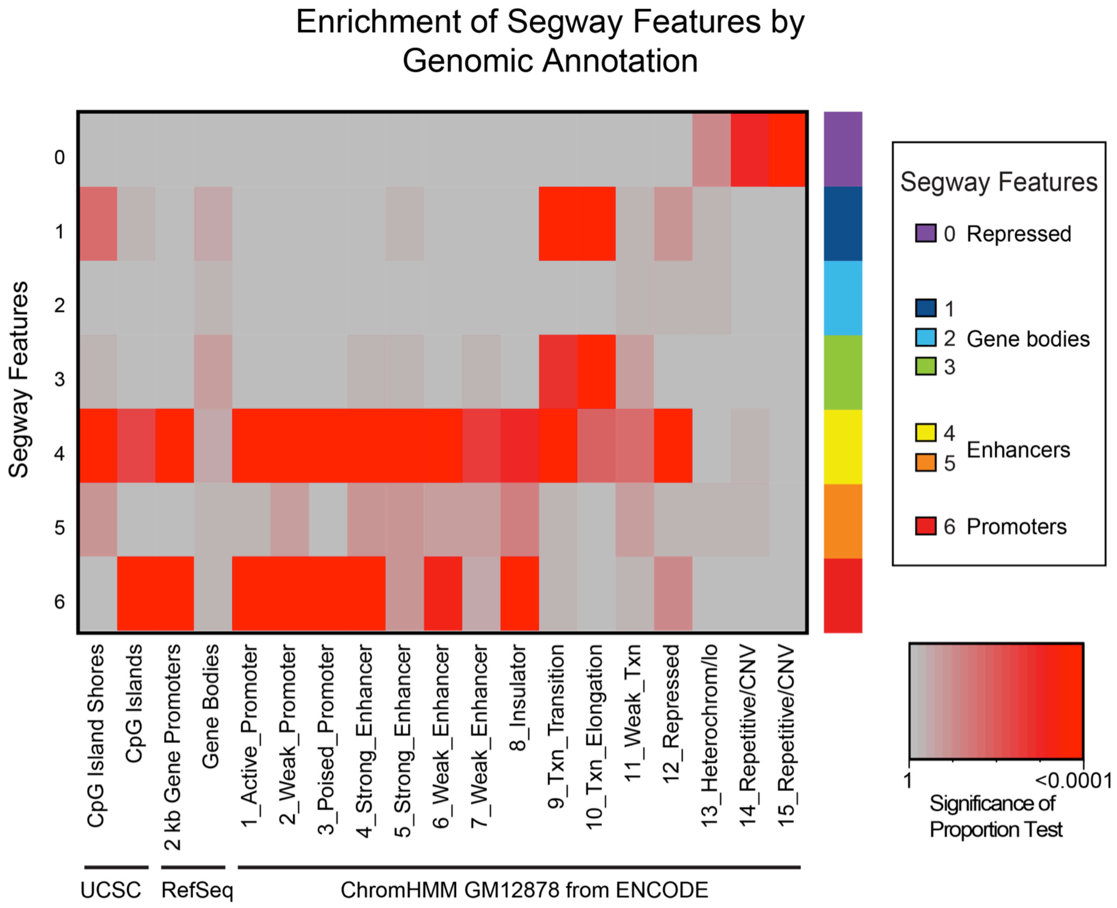

FIGURE S8: Significance testing of enrichment of genomic features and annotations for each Segway-defined feature.

ChromHMM SOM overlay

For each of the 17 ChromHMM annotation tracks, we assessed whether or not the 1,520,684 HpaII sites containing 100 bp intervals had an overlap with a ChromHMM track and created a binary indicator. Vectors were unit length normalized, and a 100 unit by 100 unit SOM was constructed using the default learnrate=0.7 and randomSeed=11. We again chose 3,080,570 iterations to keep our analyses consistent. As with the Roadmap chromatin SOM, the output wgt file was parsed to understand the local influences of ChromHMM annotations within the map and depicted as a weighted color (yellow to red). The output unit file was parsed to understand enrichment of variability within the map dichotomizing MAD, again using a cutoff of 23.09, and conducting proportion tests (alpha=0.05, **Figure S9**).

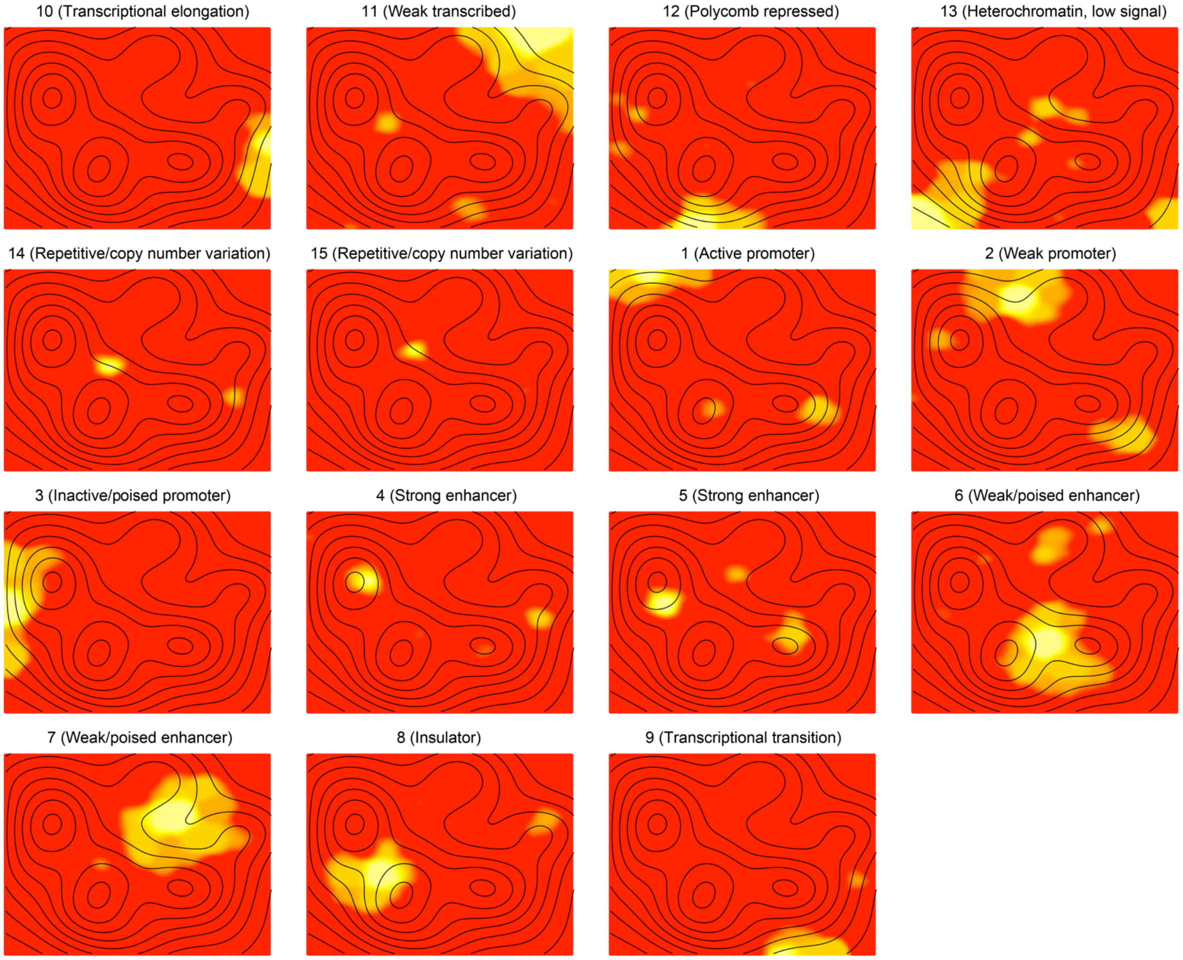

FIGURE S9: SOM analysis of ChromHMM states, with overlying contour plots for DNA methylation variability. The strongest associations appear to be states 4/5 (Strong enhancer).

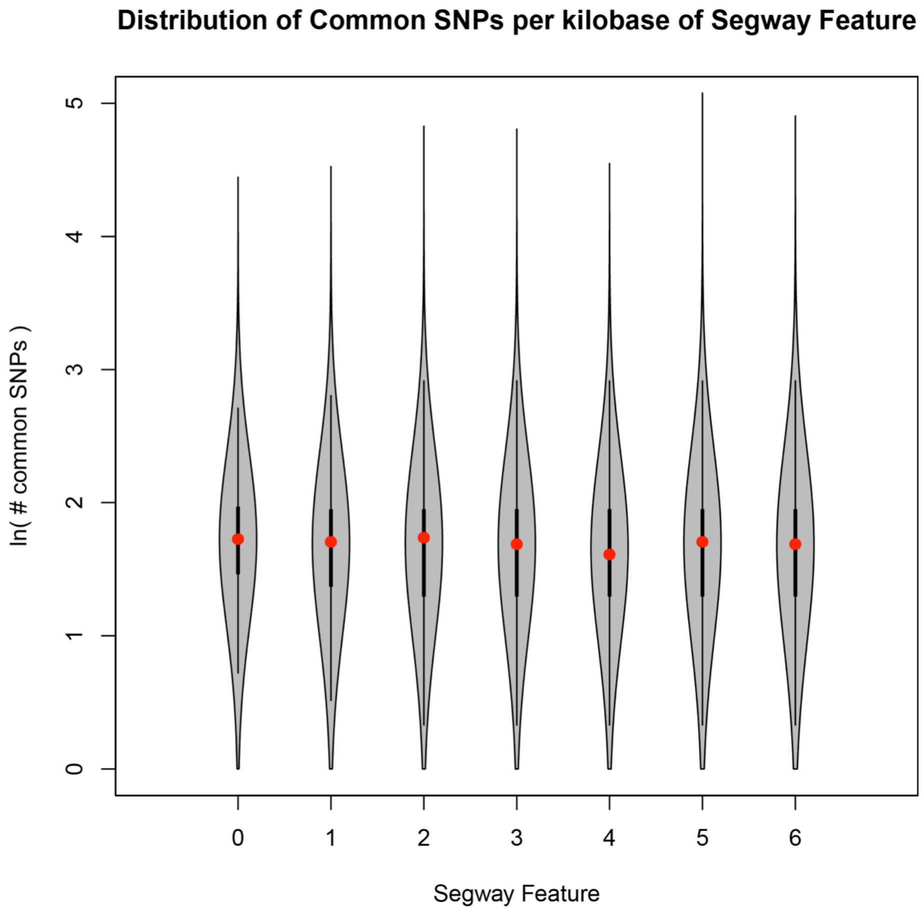

FIGURE S10: The frequency of Common SNPs was calculated per kilobase for each of the Segway-defined features. No difference in frequencies was found, indicating that differences in epigenetic variability distinguishing each feature are not solely due to the enrichment for genetic variation within a subset of features.

Functional enrichment analysis

For each RefSeq gene, Segway feature 6 (promoters) overlapping the TSS and Segway feature 4 (enhancers) occurring within 5 kb flanking the TSS were isolated. We calculated the median MAD over these features. The MAD over promoters and enhancers was dichotomized using the 23.09 value (98.5th percentile) allowing genes to be characterized as having high and low variability over both promoters and enhancers. The Broad Institute's Gene Set Enrichment Analysis web applet (http://www.broadinstitute.org/gsea/) performs a hypergeometric/Fisher's exact test on gene list supplied from the Molecular Signatures Database v3.1 (http://www.broadinstitute.org/gsea/msigdb/), to identify pathways differentially enriched for high and low variability enhancers using an False Discovery Rate (FDR) q value of less than 0.05 (**Table S6**). We further demonstrate this association using Reactome pathways isolated through the Pathway Commons (http://www.reactome.org/static_wordpress/about/ and http://www.pathwaycommons.org). Gene pathways were visualized in Cytoscape v2.8.3 with edges representing the physical interactions between nodes stored in the GeneMANIA v3.2 plugin. DNA methylation variability was high for promoters for both housekeeping (KEGG Ribosome) and leukocyte-specific (Leukocyte Transendothelial Migration) genes at candidate promoters (Segway feature 6), but substantially decreased or absent at housekeeping and not leukocyte-specific genes at candidate enhancer loci (Segway feature 4, **Figure S11**).

TABLE S6: Functional enrichment analysis

Feature 4

	Cannonical Pathway	gene in pathway	Description	genes in overlap	FDR q value
HIGH MAD n=339	REACTOME REGULATION OF IFNG SIGNALING	14	Genes involved in Regulation of IFNG signaling	4	1.61E-03
	REACTOME PRE NOTCH PROCESSING IN GOLGI	16	Genes involved in Pre-NOTCH Processing in Golgi	4	2.75E-03
	BIOCARTAGLYCOLYSIS PATHWAY	10	Glycolysis Pathway	3	5.59E-03
	KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	118	Leukocyte transendothelial migration	11	5.78E-03
	PID S1P S1P1 PATHWAY	21	S1P1 pathway	4	7.76E-03
LOW MAD n=3089	MIPS RIBOSOME CYTOPLASMIC	81	Ribosome cytoplasmic	53	2.55E-08
	KEGG RIBOSOME	88	Ribosome	55	1.45E-07
	REACTOME REGULATION OF ORNITHINE DECARBOXYLASE ODC	49	Genes involved in Regulation of ornithinr decarboxylase (ODC)	33	4.35E-06
	MIPS PA700 20S PA28 COMPLEX	36	PA700-20S-PA28 complex	26	6.34E-06
	REACTOME DESTABILIZATION OF MRNA BY AUF1 HRNP DO	53	Genes involved in Destabilization of mRNA by AUF1 (hnRNP)	34	1.56E-05

Feature 6

	Cannonical Pathway	gene in pathway	Description	genes in overlap	FDR q value
HIGH MAD n=195	MIPS 55S RIBOSOME MITOCHONDRIAL	78	55S ribosome mitochondrial	6	7.59E-03
	PID HF1APATHWAY	19	Hypoxic and oxygen homeostasis regulation f HIF-1-alpha	3	8.04E-03
	MIPS RIBOSOME CYTOPLASMIC	81	Ribosome cytoplasmic	6	9.08E-03
	KEGG RIBOSOME	88	Ribosome	6	1.34E-02
	REACTOME REGULATION OF HYPOXIA INDUCIBLE FACTOR HIF BY OXYGEN	25	Genes involved in Regulation of Hypoxia-inducible Factor (HIF)	3	1.73E-02
LOW MAD n=1281	MIPS RIBOSOME CYTOPLASMIC	81	Ribosome cytoplasmic	67	1.04E-11
	KEGG RIBOSOME	88	Ribosome	70	1.20E-10
	MIPS NOP56P ASSOCIATED PRE RRNA COMPLEX	104	Nop56p-associated pre-rRNA complex	77	6.11E-09
	MIPS 60S RIBOSOMAL SUBUNIT CYTOPLASMIC	47	60S ribosomal subunit cytoplasmic	40	3.41E-08
	REACTOME REGULATION OF MRNA STABILITY BY PROTEINS THAT BIND AU RICH ELEMENTS	84	Regulation of mRNA Stability by Proteins that Bind AU-rich Elements	62	2.25E-07

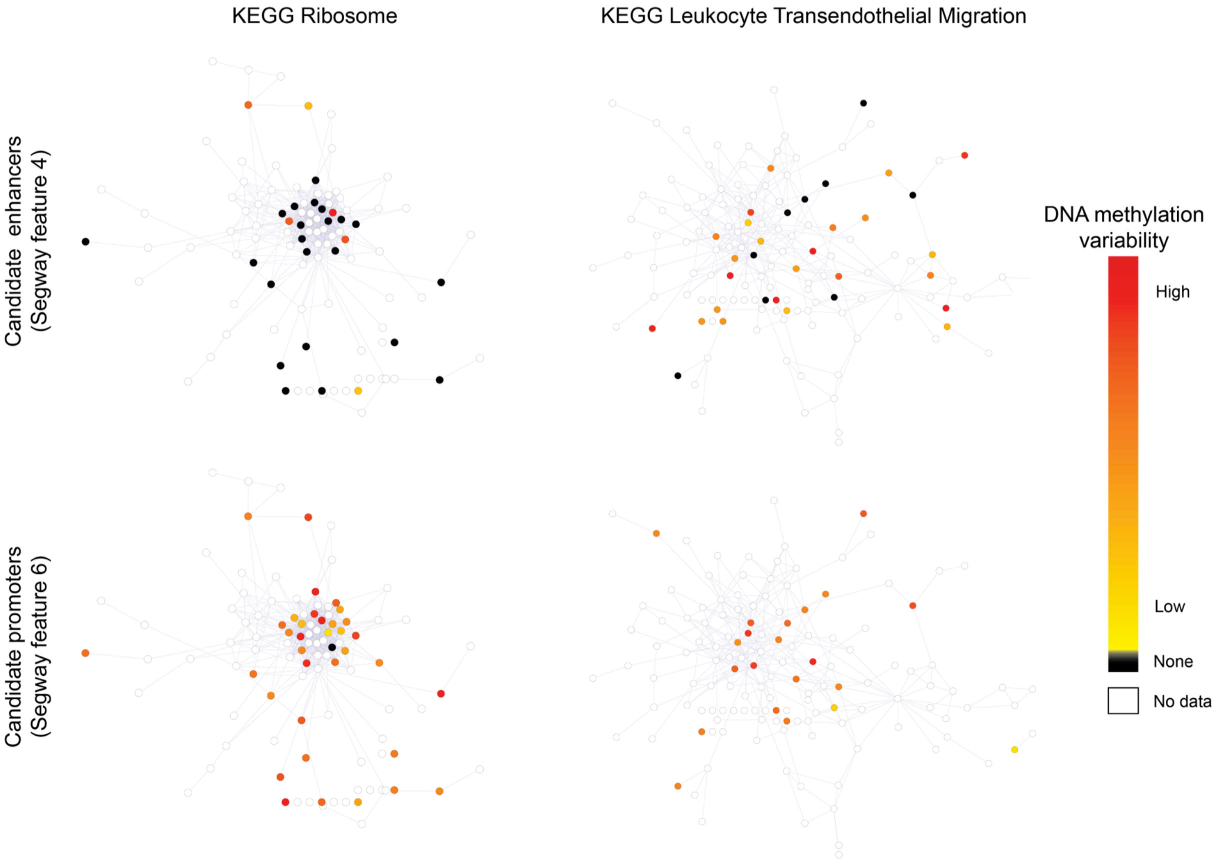

FIGURE S11: DNA methylation variability is illustrated for housekeeping genes (KEGG Ribosome network, left) and leukocyte-specific genes (KEGG Leukocyte transendothelial migration network, right). While the DNA methylation variability at Segway feature 6 (candidate promoter) annotations near these genes is high for both networks, the substantially decreased variability for the candidate enhancer loci (Segway feature 4) near the housekeeping genes is clear. The DNA methylation variability at promoters does not distinguish between these types of gene groups as effectively as the enhancer sequences.

Non-negative matrix factorization (NMF)

Application to DNA methylation data

NMF has been employed successfully in deconvolving gene expression data [13], but it has yet to be applied to DNA methylation datasets. The goal of an NMF algorithm is to deconvolve a matrix *V*, a (n x p) matrix, in order to find an approximation the matrices *W* and *H* such that:

$$V \approx WH,$$

where W and H are (n x r) and (r x p) non-negative matrices. The rank of the matrix (r) should be greater than 0 and at least 2 to represent 2 subpopulations. Because the methylation outcome is binary, and the existing NMF algorithms allow DNA methylation to vary without this constraint the estimated matrix, W' may not be interpretable directly, and improvement of the technique may allow the methylation pattern of the individual subpopulations to be estimated. We applied an existing NMF algorithm to understand the presence of subpopulations within our dataset but did not interpret the specific values within W', rather we focused on the difference between the actual and simulated datasets.

Calculation of cell subtype number

Utilizing the R package *deconf*, we varied the matrix rank and estimated the matrices W' and H' such that:

$$V' \approx W'H'.$$

The distance between the original dataset V and V' was calculated as the Frobenius norm:

$$||V' - V||_F^2 = F.$$

The process was repeated 100 times per value of r, subsetting the data to 10,000 HpaII sites in order to make the algorithm computationally tractable. A plot of the distribution of *F* as a function of increasing r (cell subpopulations) is shown (**Figure 6**). In addition, the distribution of Frobenius norms for each cell subpopulation was compared to the preceding cell subpopulation with a two sample t-test, testing for a difference in distribution of Frobenius norms between successive simulation levels. Because estimating additional subpopulations will always explain additional variability, a smooth spline was fit to the data to look for inflection points indicating a local minimum in the Frobenius norm when related to cell subpopulations.

SUPPLEMENTARY REFERENCES

1. Landmann, E., et al., *Ponderal index for discrimination between symmetric and asymmetric growth restriction: percentiles for neonates from 30 weeks to 43 weeks of gestation.* J Matern Fetal Neonatal Med, 2006. **19**(3): p. 157-60.
2. Pranke, P., et al., *Hematologic and immunophenotypic characterization of human umbilical cord blood.* Acta Haematol, 2001. **105**(2): p. 71-6.
3. Willasch, A., et al., *Enrichment of cell subpopulations applying automated MACS technique: purity, recovery and applicability for PCR-based chimerism analysis.* Bone Marrow Transplant, 2010. **45**(1): p. 181-9.
4. Verma, A., et al., *Activation of the p38 mitogen-activated protein kinase mediates the suppressive effects of type I interferons and transforming growth factor-beta on normal hematopoiesis.* J Biol Chem, 2002. **277**(10): p. 7726-35.
5. Suzuki, M., et al., *Optimized design and data analysis of tag-based cytosine methylation assays.* Genome Biol, 2010. **11**(4): p. R36.
6. MacDonald, J.W. and D. Ghosh, *COPA--cancer outlier profile analysis.* Bioinformatics, 2006. **22**(23): p. 2950-1.
7. Jaffe, A.E., et al., *Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.* Int J Epidemiol, 2012. **41**(1): p. 200-9.
8. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC Bioinformatics, 2009. **10**: p. 232.
9. Zhu, Q., et al., *A novel recursive Bayesian learning-based method for the efficient and accurate segmentation of video with dynamic background.* IEEE Trans Image Process, 2012. **21**(9): p. 3865-76.
10. *Self-Organizing Maps*, ed. T. Kohonen, M.R. Schroeder, and T.S. Huang. 2001: Springer-Verlag New York, Inc. 528.
11. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization.* Nat Methods, 2012. **9**(3): p. 215-6.
12. Hoffman, M.M., et al., *Unsupervised pattern discovery in human chromatin structure through genomic segmentation.* Nat Methods, 2012. **9**(5): p. 473-6.
13. Gaujoux, R. and C. Seoighe, *Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study.* Infect Genet Evol, 2012. **12**(5): p. 913-21.